Computational Linguistics in the Netherlands Journal 14 (2025) 43-65 Submitted 12/2024; Published 07/2025

Evaluating LLM-Generated Topic Names via Text Reconstruction

Andriy Kosar[*],[**] ANDREW@TEXTGAIN.COM
Mariia Butynets[*] MARY@TEXTGAIN.COM
Guy De Pauw[*] GUY@TEXTGAIN.COM
Walter Daelemans[**] WALTER.DAELEMANS@UANTWERPEN.BE

[*] *Textgain, Antwerp, Belgium*

[**] *University of Antwerp (CLiPS), Antwerp, Belgium*

Abstract

Automatically generating topic names for texts using large language models (LLMs) has become an innovative approach to topic detection. However, evaluating the quality of these LLM-generated topic names remains challenging, particularly in assessing their semantic relevance to the texts and the correctness of the information they convey. To address this gap, we propose a novel evaluation method that leverages LLMs to reconstruct original texts from generated topic names, then compares the reconstructed texts to the original by measuring their similarity. Topic names that produce reconstructed texts more similar to original ones better convey the original text's information. This method favors topic names that maintain essential information, minimizing issues like incorrectness and irrelevance. Our experiments show that the reconstruction-based evaluation aligns with human topic name evaluation. This novel method demonstrates its versatility for evaluating other LLM-generated semantic compressions, such as summaries, headlines, and keywords.

1. Introduction

Topic detection plays an essential role in organizing and interpreting large volumes of digital content, enhancing the accessibility and utility of information across various fields. For the last two decades, the most common method for topic detection has been probabilistic topic modeling. For example, a typical topic generated by this method might be represented by the top words “machine”, “learning”, “neural”, “networks”, and “transformers”. As outlined by Boyd-Graber et al. (2017), topic modeling has been instrumental in applications such as news categorization and general information retrieval. However, with recent advancements in LLMs, the scope of topic detection has expanded far beyond traditional uses and methods. Today, LLMs are applied to generate topic names across diverse formats, including short names for chat conversations (e.g., in ChatGPT or Cohere chats), naming email threads, labeling podcasts based on discussed topics, naming groups of social media comments, and automatically naming folders for groups of documents. These applications support improved content organization and more efficient categorization and retrieval of information. For instance, an LLM-generated topic name might look like “Deep Learning Innovations: Transformer Advances”.

The emergence of LLMs has enabled significant advancements in automatically generating topic names, with these models now performing at near-human levels (Kosar et al. 2024). However, evaluating the quality of these generated topic names remains unexplored. It is difficult to establish a consistent benchmark labeled by humans for evaluating automatically generated topics, as human-named topics can vary widely depending on background and subjective interpretation. Also, applying existing evaluation methods from probabilistic topic modeling to LLMs-generated topic names is problematic because of the differences in how topics are defined and represented in these approaches.

©2025 Andriy Kosar, Mariia Butynets, Guy De Pauw, and Walter Daelemans.

In Section 2.1, we discuss in detail why existing evaluation methods are not applicable for evaluating LLMs-generated topic names.

To address these challenges, this work introduces a novel evaluation method for generated topic names based on the principle that a meaningful topic name should effectively convey the essential information in the text. Our method uses LLMs to reconstruct the original text from a generated topic and then evaluates the similarity between the reconstructed text and the original. This reconstruction-based method emphasizes information preservation, counteracting cases where the generated topic may include extraneous or misleading information. Metrics such as ROUGE (Lin 2004), BERTScore (Zhang et al. 2020), and cosine similarity (Wang and Dong 2020) are used to quantify the alignment between the original and reconstructed texts, favoring topics that preserve essential information. We demonstrate that this reconstruction-based evaluation is broadly applicable to other LLM-generated compressions, including summaries, headlines, and keywords, highlighting its versatility as a method for assessing condensed representations of text.

The primary contributions of this paper are as follows:

1. Introduction of a reconstruction-based method for topic name evaluation. We present a novel evaluation method that leverages LLMs to reconstruct the original text from generated topics, focusing on information preservation to enable an automatic assessment of topic quality.
2. Validation against human judgment. We validate our method by comparing it with human evaluations, establishing its reliability and alignment with human perspectives on topic quality.
3. Applicability to other forms of semantic compression. Our method extends beyond topic names to evaluate other types of LLM-generated condensed representations, such as summaries, headlines, and keywords, demonstrating its adaptability and potential for broader applications.

The structure of the paper unfolds as follows: Section 2 provides context for our research by discussing topic modeling evaluation metrics, recent developments in applying LLMs for topic detection, and related studies on semantic compression. Section 3 highlights specific studies in linguistics, information theory, and communication theory that have influenced the proposed method, defines the problem, formulates the hypotheses, and offers an overview of the proposed method. Section 4 describes the experimental setup, including data preparation, and the selected evaluation approaches and procedures for the proposed method. Section 5 presents the findings of the intrinsic evaluation and validates the proposed method on human preferences. Section 6 outlines the limitations of the study. Finally, Section 7 summarizes the experimental outcomes and suggests directions for future research.

2. Related Work

2.1 Topic Modeling Evaluation Metrics

In topic modeling, evaluation metrics such as perplexity (Blei et al. 2003), coherence (Newman et al. 2010, Mimno et al. 2011), diversity (Dieng et al. 2020), coverage (Korencic et al. 2021), and significance (AlSumait et al. 2009) capture various aspects of model quality. For a comprehensive overview of these metrics, see Churchill and Singh (2022) and Hoyle et al. (2021). To further validate these metrics, researchers have conducted studies incorporating human judgment, often introducing word intruders into topic keywords or ranking keywords by relevance (Chang et al. 2009, Newman et al. 2010, Lau et al. 2014, Bhatia et al. 2017, Lund et al. 2019, Hoyle et al. 2021). Recent research suggests that LLMs can enhance the usage of these metrics for topic model evaluation by leveraging their pre-trained semantic understanding. For instance, Stammbach et al. (2023) recommend employing LLMs to evaluate topic coherence, as LLMs' scores correlate strongly with human judgment. Moreover, LLM-based evaluation is receiving increased academic attention in general, as noted in Li et al. (2024), which provides a comprehensive overview and discusses its limitations.

However, these metrics are not directly applicable to evaluating LLM-generated topic names, as they are primarily designed to assess model fit and the internal qualities of the model rather than the quality of topic names, particularly in terms of how well they capture and convey the main information about the text. For example, *perplexity* measures how well a model fits held-out data, but it does not evaluate the quality of topic names. *Coherence* evaluates the semantic consistency among a set of topic keywords based on word co-occurrence in target data, yet it does not measure whether the resulting topic name is meaningful or related to the text. Similarly, *diversity* assesses the variety of topic names by counting unique words among all topics, and *coverage* assesses how well a topic model captures a set of predefined concepts, but neither metric addresses how effectively the name describes the underlying content. Consequently, none of these metrics address how well the topic name conveys the main information about the text, how accurate that information is, or how clearly the topic name is presented to the reader.

2.2 LLM Capabilities and Topic Detection

The advent of LLMs has enabled significant advancements in topic detection and topic naming. Kosar et al. (2024) investigated LLMs' ability to detect and label topics in news texts, comparing LLM performance to human-generated topics. Through experiments, they found that LLMs could generate topic names comparable to those produced by humans, though individual differences in topic perception and naming persist. Kirilenko and Stepchenkova (2024) compared LLMs and Latent Dirichlet Allocation (LDA) for analyzing social media reactions, specifically in the tourism domain. The authors demonstrated that LLMs outperform LDA in handling short, noisy texts, suggesting LLMs' effectiveness for topic modeling in similar contexts.

Furthermore, Kapoor et al. (2024) introduced a method that applies LLMs to extract key phrases from texts, subsequently clustering and subclustering these phrases, and then using LLMs to identify the main themes for each cluster. Their approach demonstrated enhanced performance on both coherence and diversity benchmarks. Mu et al. (2024) proposed using LLMs for topic modeling, overcoming the limitations of traditional methods like LDA. Their approach prompted LLMs to generate coherent, human-like topic titles without requiring additional training, demonstrating that LLMs produce relevant and interpretable topics for large text corpora.

2.3 Semantic Compression and Information Preservation

In addition to topic generation, researchers have explored methods for semantic compression to represent textual information efficiently. Gilbert et al. (2023) proposed using LLMs for semantic compression by exploring various prompting techniques to guide the models in generating compressed representations of text that preserve meaning. They also introduced the Semantic Reconstruction Effectiveness (SRE) metric, which uses cosine similarity between the embeddings of the original and decompressed text to evaluate the effectiveness of the compression approaches.

Yu and Varshney (2024), drawing on Claude Shannon's "lattice theory of information", proposed applying this framework to represent information hierarchically, with varying levels of abstraction. This approach facilitates compression by simplifying data while preserving essential meaning at different semantic levels. The authors introduced a technique called "information lattice learning" to uncover these hierarchical semantic structures within data and leverage them for compression.

3. Methodology

3.1 Problem Definition

Evaluating topic names, whether authored by humans or systems, is inherently complex, as it requires assessing multiple aspects, including relatedness, completeness, clarity, and correctness, along with other aspects that are discussed in detail in Kosar et al. (2024). With the emergence of LLMs in topic generation, new challenges such as "model hallucination" (Huang et al. 2024), where models may generate irrelevant or incorrect information, have become more pronounced. These challenges highlight the need for evaluation methods that effectively address these aspects in assessing generated topic names.

We propose that an effective topic name should efficiently communicate the essential information ("aboutness") about the topic of the text. This aligns with Zipf's principle of least effort (Zipf 1949), which suggests that communicators naturally aim to convey information as efficiently as possible, minimizing cognitive load for both speaker and listener. Additionally, Grice's maxims of conversation (Grice 1975) emphasize that effective communication balances informativeness with clarity, relevance, and brevity. Therefore, our evaluation focuses on how much of the text's information is captured by the topic name and how accurately this information can be restored from it by humans or systems. This assessment accommodates important aspects of topic name evaluation: relevance, completeness, clarity, and correctness. By ensuring that the topic name enables more lossless text reconstruction, we ensure that it aligns with the main topic of the text (relevance), captures essential information (completeness), reduces ambiguity (clarity), and maintains factual accuracy (correctness).

Thus, we formulate our hypotheses as follows:

1. A topic name's quality can be quantitatively assessed by the degree of information preservation it enables, with high-quality topic names preserving more of the source text's essential information.
2. Users tend to prefer topic names that retain a higher level of information from the source.

3.2 Proposed Method

The proposed method evaluates topic names, whether generated by LLMs or written by humans, by examining their ability to preserve and convey essential information about the described topic. This approach draws from information theory and communication theory, particularly focusing on the processes of encoding and decoding information as conceptualized in the Shannon-Weaver Model of Communication (Shannon 1948).

In communication theory, a sender encodes original information into a message, which is then transmitted to a receiver who decodes it to reconstruct the original information based on their knowledge and understanding. Analogously, in our method, the original text serves as the primary information source, while topic names, whether system-generated or human-constructed, act as a compressed form of the original text's semantics, functioning as the encoded message. To emulate the human ability to decode and understand information from a topic name, we utilize LLMs. Given their proficiency in interpreting and generating text, LLMs are well-suited for reconstructing the original content from topic names. This process mirrors Bartlett's Schema Theory (Bartlett 1932), which posits that individuals interpret and reconstruct information based on mental frameworks or schemas formed from prior experiences.

This method encompasses both the reconstruction process and the comparison of the original text with the reconstructed text. We propose using cosine similarity as the primary metric for this comparison, measuring the semantic similarity between the two texts. A high cosine similarity score indicates that the topic name effectively preserves and conveys the core information of the original text. While acknowledging that some information loss is inevitable during compression, especially with topics, the comparison between the original and reconstructed texts provides an approximation of information preservation.

This reconstruction-based evaluation method, as described in Figure 1, allows us to automatically assess the quality of topic names in terms of their information preservation, as well as compare topic models by evaluating generated topics. Our approach is adaptable to a range of LLM architectures and can be applied across various models, from general-purpose LLMs to fine-tuned versions. Furthermore, this method is applicable to other forms of text compression, such as summaries, headlines, and keywords, providing a versatile method for evaluating various condensed representations of textual information.

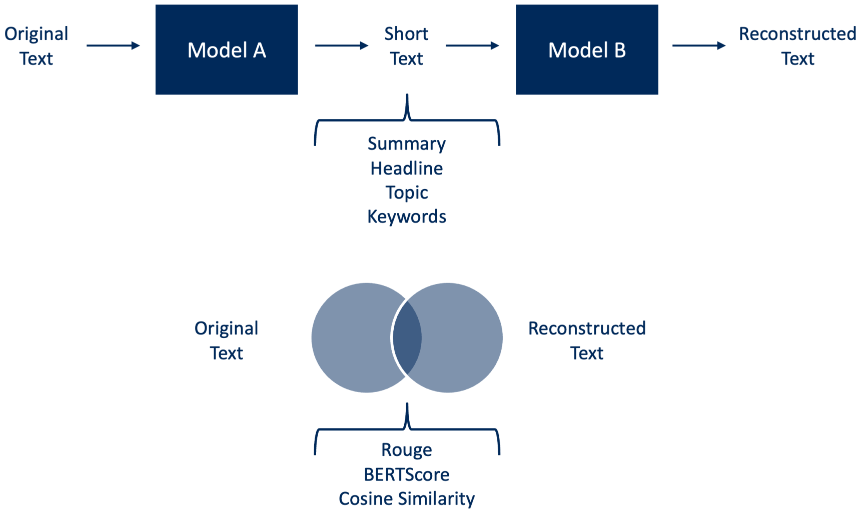

Figure 1: Topic name evaluation framework through text reconstruction.

4. Experiment Setup

4.1 Text Selection

For our study, we utilized 100 articles written in English from The Guardian, published in the first quarter of 2024. We selected 10 articles randomly from each of the 10 different sections of the publication, as listed in Appendix A.2.4. The motivation behind this was to utilize recent content, minimizing the likelihood of overlap with training data used in existing LLMs. This approach allowed us to evaluate our method on contemporary, diverse texts.

4.2 Topic Generation

For each of the 100 texts, we generated topic names using four models, resulting in a total of 400 topic names. The models used were GPT-4 (OpenAI 2023)[1], GPT-3.5 (Ouyang et al. 2022)[2], GPT-3.5 Fine-tuned (GPT3.5FT), and FLAN-T5 Fine-tuned (FLAN-T5FT). The GPT-4 and GPT-3.5 models were provided with examples of both high-quality and low-quality topic names from the study conducted by Kosar et al. (2024).

---

[1] Model version: gpt-4-1106-preview

[2] Model version: gpt-3.5-turbo-0125

The GPT-3.5FT model[3] was fine-tuned using 1,000 topics generated by GPT-4 using the Newsroom dataset (Grusky et al. 2018). Similarly, we fine-tuned the FLAN-T5 model (Chung et al. 2022)[4] on 20,000 text and topic name pairs from the Newsroom dataset, previously generated by the GPT-3.5FT. Examples of generated topics are provided in Appendix A.2.4. Our motivation was to accommodate a variety of topic names and to explore whether these smaller, fine-tuned models could achieve performance comparable to their larger counterparts.

4.3 Other Compressed Texts

In addition to evaluating topic names, our study also explored other forms of semantic compression, such as headlines, summaries, and keywords. For the headlines, we used the original ones authored by news article writers. We generated three summaries for each news article, yielding a total of 300 summaries. For this, we employed the following models: GPT-4, GPT-3.5, and a fine-tuned T5 model (T5FT) (Raffel et al. 2020). Keywords, which often serve as concise representations of topics, were generated using the KeyBERT model (Grootendorst 2020)[5]. We assessed sets of keywords in various sizes 5, 7, and 10 to determine how well they captured the main themes and details of the articles. Table 1 summarizes the token distribution for each type of compression, except for keywords, which have fixed length. Examples of compressed texts and their reconstructions are provided for illustrative purposes in Appendix A.2.2 and A.2.3.

Method/Model	**Tok mean**	**Tok median**	**Tok max**	**Tok min**	**Tok std**	**N records**
Summaries GPT4	24.4	24.0	44	17	4.0	100
Summaries GPT3.5	19.4	20.5	39	5	8.6	100
Summaries T5FT	18.4	18.0	30	12	3.1	100
Headlines	13.9	14.0	24	9	3.0	100
Topics GPT4	5.3	5.0	9	3	1.1	100
Topics GPT3.5	5.2	5.0	8	3	1.1	100
Topics GPT3.5FT	5.5	5.0	8	4	1.2	100
Topics FLANT5FT	5.4	5.0	8	3	1.1	100

Table 1: Dataset description for different compression methods and models.

4.4 Text Reconstruction and Similarity Evaluation

For the task of reconstructing the original text from the generated topic names, keywords, headlines, and summaries, we utilized the following three models: GPT-4, GPT-3.5, and Mixtral-8x7B (Jiang et al. 2024)[6]. All models were given the same prompt to reconstruct news text (see Appendix A.2.1). As a result, we obtained 1,200 reconstructed texts using topic names, 900 using keywords, 300 using headlines, and 900 using summaries.

To assess information preservation in compressed texts, we measured the similarity between the original and reconstructed texts using three primary metrics: ROUGE, BERTScore, and cosine similarity. These metrics quantified how well the reconstructed text retained the essential information and meaning of the original text through different aspects: ROUGE for measuring lexical overlap using n-gram and word sequence comparisons, BERTScore for assessing semantic similarity of contextual embeddings of words in context, and cosine similarity for evaluating the overall semantic similarity using vector representations of the texts. Each of these metrics has its own limitations (e.g., see Zhang et al. (2024) for ROUGE, Hanna and Bojar (2021) for BERTScore, and Zhou et al. (2022) for cosine similarity), and they serve as approximate indicators of similarity.

---

[3] Base model version: gpt-3.5-turbo-1106

[4] Base model version: https://huggingface.co/google/flan-t5-base (Accessed January 27, 2024)

[5] We used KeyBERT in its default configuration: https://github.com/MaartenGr/KeyBERT (Accessed February 13, 2024)

[6] The model was accessed via the https://www.together.ai/ API (Accessed February 15, 2024)

We employed the BAAI General Embedding (BGE) model (Xiao et al. 2024)[7] to generate text embeddings for both the original and reconstructed texts, which were then used to calculate cosine similarity.

4.5 Intrinsic Evaluation

Our intrinsic evaluation of the proposed method focused on the following aspects: preservation of information across different compression methods, the impact of various models on information preservation in summaries and topics, the effectiveness of reconstruction models, the utility of existing text similarity metrics, and the consistency of results across these different dimensions.

1. **Compression Methods**: We analyzed how well different text compression methods, namely headlines, summaries, and keywords, preserve and convey essential information from the original texts. By comparing reconstructed texts with the original articles, we assessed the effectiveness of each method in maintaining and conveying the main information about the text.
2. **Compression Models**: We evaluated the effect of various models (4.2, 4.3) on the preservation of information in summaries and topics. This analysis aimed to assess whether the models had a consistent impact across different methods (summaries and topics) and within each method.
3. **Reconstruction Models**: We assessed the impact of reconstruction models on information preservation (4.4) across various compression methods. This analysis aimed to examine whether reconstructing texts using different LLM models changes the effect of compression methods and models on the preservation of information in compressed texts (e.g., topics, summaries, etc.).
4. **Text Similarity Metrics**: We evaluated the utility of various similarity metrics (4.4), assessing how suitable they are for measuring the differences between original and reconstructed text and how well they fit the purpose.

This comprehensive assessment aimed to thoroughly examine the proposed evaluation method, its capabilities and limitations, and to assess the effect of various models on preserving information across different forms of text compression.

4.6 Validation on Human Preference Data

We also validated our method against human judgment drawn from existing studies. We used existing human preference studies because it was infeasible to conduct a new human preference study within the scope of this paper for our dataset of 1200 topic names and 900 summaries. To validate the proposed method, we followed these steps: reconstructing texts from existing compressed texts (topics and summaries), measuring similarity between the original and reconstructed texts, and conducting a correlation analysis between text similarity and human judgment, using Spearman's rank correlation. We applied the Qwen2 model (Yang et al. 2024)[8] to reconstruct texts from topics and summaries. Since we used an OpenAI dataset (summaries), we chose to employ a non-OpenAI model for data reconstruction to avoid cases where possible data contamination could lead to biased results. To generate text embeddings for both the original and reconstructed texts, we employed the BGE model.

For topics, we referenced the human evaluation study conducted by Kosar et al. (2024). In their research, four judges evaluated 96 topics, encompassing both human-written and LLM-generated ones. The human evaluation score for each topic was calculated by averaging judges' scores on a scale from 1 to 3, with higher scores indicating better quality. We reconstructed texts from the

---

[7] Model version: BAAI/bge-base-en-v1.5 – https://huggingface.co/BAAI/bge-base-en-v1.5 (Accessed June 12, 2024)

[8] Model version: Qwen2-72B-Instruct, accessed via the https://www.together.ai/ API (Accessed June 12, 2024)

topics used in the study and then measured the similarity between the reconstructed texts and the original news texts for which topic names were assigned. By measuring the correlation between the text similarity scores and the human evaluation scores, we aimed to determine how well they align.

For summaries, we leveraged the OpenAI human preferences research on improving summary quality by training models to optimize for human preferences, as detailed by Stiennon et al. (2022). The original study generated summaries using GPT-3 models with 1.3 and 6.7 billion parameters, trained with different reward policies, and evaluated these summaries based on human preferences across four dimensions: coverage, accuracy, coherence, and overall quality, using a 7-point Likert scale. In this context, as defined by the authors, “coverage” refers to the extent to which the summary captures the key information from the text, “accuracy” to the degree to which the statements in the summary are stated in the text, “coherence” to how easy the summary is to read, and “overall” to how good the summary is overall at representing the text. In our study, we used summaries from the test split for articles from the Daily Mail and CNN with more than 30 characters, resulting in 5,552 summaries for 639 news texts. After reconstructing texts from summaries, we conducted a correlation analysis between the text similarity score and each of the four dimensions used in human evaluation.

5. Results

5.1 Intrinsic Evaluation

Compression Methods

Our analysis of different compression methods with regard to information preservation showed that there was a clear difference between the compression methods. The highest values in text similarity scores were consistently obtained by summaries, followed by headlines, topics, and keywords, according to all metrics: ROUGE-1 (F1), BERTScore (F1), and cosine similarity. This pattern was observed for the entire dataset (Figure 2) and also when analyzed by reconstruction models.

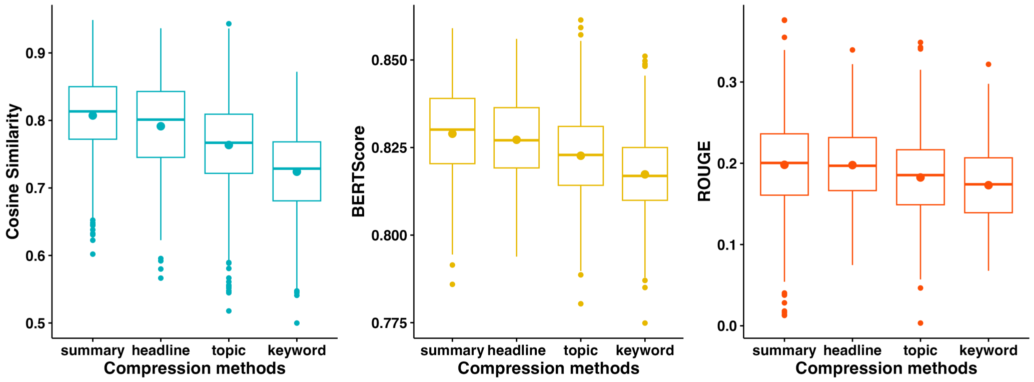

Figure 2: Side-by-side box plots of text similarity scores for four compression methods: 1200 topics, 900 keywords, 300 headlines, and 900 summaries. The text similarity metrics are: cosine similarity, BERTScore F1, and ROUGE-1 F1.

We observed that compression methods resulting in longer texts generally yield better reconstruction of the original text due to the preservation of more information, except in the case of keywords. In the case of keywords, we found that this method preserves less information than topics, even when using the same number or more tokens. We hypothesize that this is due to the method’s inherent limitations – the absence of syntax – which creates ambiguity in how relationships between keywords can be interpreted, thus diluting the semantics of the compression.

Reconstruction Models

The analysis of the impact of the compression methods on text similarity metrics, when conducted separately for each reconstruction model, showed a consistent downtrend in metric scores starting from summaries, followed by headlines, topics and keywords (Figure 3, and Appendix A.1, Figures 8 and 9). All compression methods were distinguishable, except for summaries and headlines for cosine similarity and BERTscore (F1) in Mixtral-8x7B, and for ROUGE-1 (F1) in GPT3.5 and Mixtral-8x7B models.

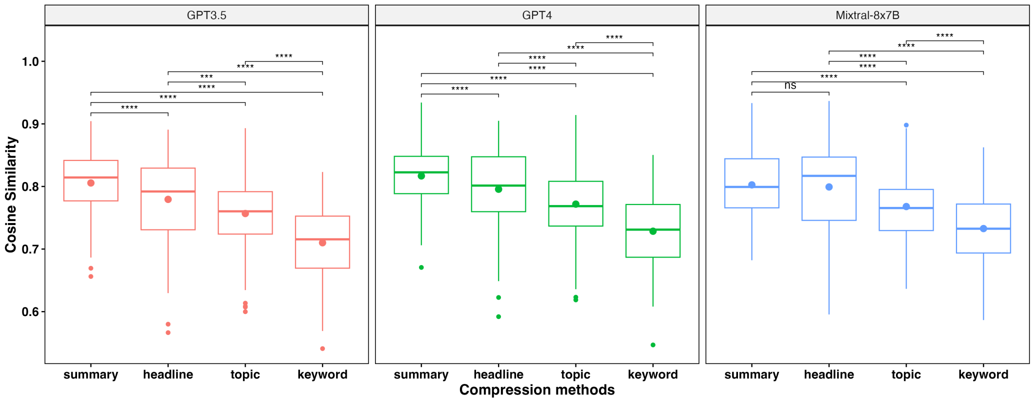

Figure 3: Side-by-side box plots of the cosine similarity in four compression methods, split by three reconstruction models. Each compression method group contains 100 records, with each record representing the median metric score for compression models within that method. Statistically significant mean differences between method pairs are denoted by asterisks. Method: two-way repeated measures ANOVA, followed by pairwise paired t-test comparisons with Benjamini–Hochberg correction.

Our analysis of the impact of the reconstruction model on the quality of text reconstruction concerning the compression models within the compression method revealed that the differences between compression models remain, no matter which reconstruction models were used (Appendix A.1, Tables 3 and 4). Based on these observations, we conclude that while the quality of the reconstruction model can affect the quality of the reconstruction based on its knowledge of the subject and proficiency in the task, we would expect better performance if more information is preserved by the compression method in the compressed text.

Text Similarity Metrics

Text similarity metrics such as ROUGE-1, BERTScore, and cosine similarity effectively differentiated compression methods, indicating that these metrics can quantify how much information is preserved when comparing original and reconstructed texts (Figure 3, Appendix A.1, Figures 8 and 9). However, their ability to distinguish between summaries and headlines is not always consistent, as in the case of texts reconstructed by the Mixtral-8x7B model.

For summaries, each metric demonstrated its capability to effectively differentiate between the compression models, showing that ROUGE-1, BERTScore, and cosine similarity can be used for evaluating summaries and, potentially, the models that generate those summaries. For topics, cosine similarity excelled at distinguishing among the different topic models, unlike ROUGE-1 and BERTScore, which did not show statistically significant differentiation. This might indicate that cosine similarity is better suited for capturing the semantic differences in topic-based text reconstructions than more token-based similarity comparisons provided by ROUGE-1 and BERTScore. In the case of keywords, we observed that all metrics differentiate between texts reconstructed from various keyword lengths (5, 7, and 10 words, respectively) and showed an uptrend in performance with the increasing number of keywords (Appendix A.1, Figure 10).

These observations indicate that the mentioned text similarity metrics can be applied to quantify information preservation when comparing original and reconstructed text while accounting for variations introduced by different compression methods and models. Moreover, they underscore that cosine similarity is suitable for topic evaluations due to its ability to capture semantic differences between texts that may have similar word distributions.

Compression Models

The evaluation of models generating summaries revealed that GPT-4 outperforms GPT-3.5 across all metrics, aligning with the general performance improvements of GPT-4 compared to GPT-3.5. GPT-4 also surpasses the T5FT model in BERTScore (F1) and ROUGE-1 (F1) while showing no significant difference in cosine similarity. Additionally, we observe that the T5FT model outperforms GPT-3.5 in cosine similarity and BERTScore (F1), with no significant difference in ROUGE-1 (F1). Figure 4 illustrates these performance differences. Overall, our findings are consistent with other studies on GPT-4's performance and highlight the enhanced capabilities of fine-tuned models compared to general LLMs.

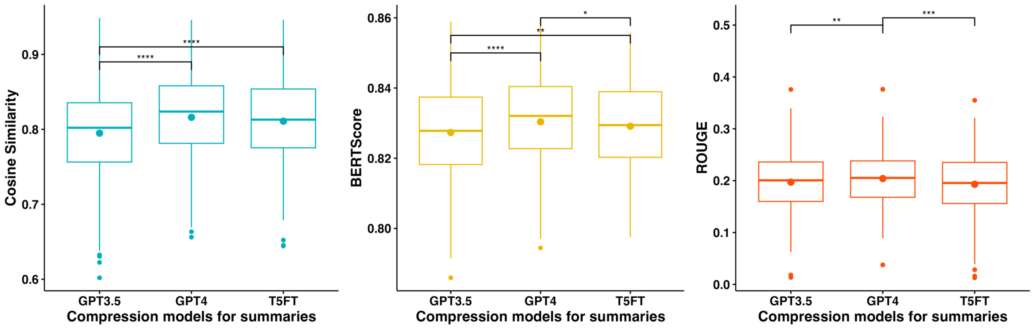

Figure 4: Side-by-side boxplots of the text similarity metrics (cosine similarity, BERTScore F1, and ROUGE-1 F1) for summaries generated by three compression models (GPT-3.5, GPT-4, and T5FT). Each compression model group contains 300 records. Statistically significant mean differences between model pairs are denoted by asterisks. Method: two-way repeated measures ANOVA, followed by pairwise paired t-test comparisons with Benjamini–Hochberg correction.

The evaluation of models used to generate topic names shows that fine-tuned models on specific tasks outperform general models, as illustrated in Figure 5. This is particularly evident in the performance of GPT-3.5FT and FLAN-T5FT compared to their more general counterparts. Statistical analysis indicates that the mean cosine similarity score of FLAN-T5FT is significantly higher than those of GPT-3.5 and GPT-4. Similarly, the mean cosine similarity of GPT-3.5FT is significantly higher compared to that of GPT-3.5, suggesting that fine-tuning enhances a model's ability to generate topic names that retain more information. Comparisons among other models also reveal that

GPT-4's mean metric is statistically higher than GPT-3.5's. Examples of topics and corresponding scores are provided in Appendix A.2.4.

Overall, while GPT-3.5FT and FLAN-T5FT tend to achieve higher cosine similarity scores on average, there is considerable variability within and between models. The differences in performance demonstrate that FLAN-T5FT and GPT-3.5FT generally outperform other models in preserving information. For the other metrics, there were no statistically significant differences in the mean scores across the models used.

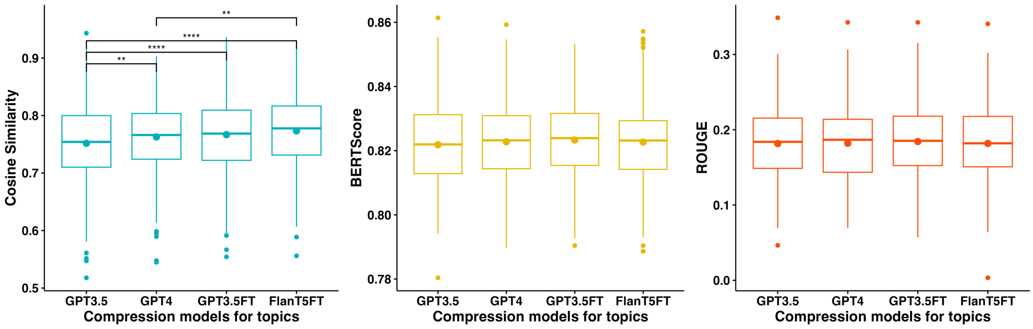

Figure 5: Side-by-side boxplots of three metric scores (cosine similarity, BERTScore F1, and ROUGE-1 F1) for topics generated by different compression models (GPT-3.5, GPT-4, GPT-3.5FT, and FLAN-T5FT). Each compression model group contains 300 records. Statistically significant mean differences between model pairs are denoted by asterisks. Method: two-way repeated measures ANOVA, followed by pairwise paired t-tests with Benjamini–Hochberg correction.

5.2 Evaluation on Human Preference Data

Topics

We found statistically significant positive monotonic associations between the human evaluation scores and similarity scores between original and reconstructed texts (Figure 6). The Spearman's correlation coefficient between the human evaluation score and cosine similarity was 0.375 with a p-value of less than 0.001. For other metrics, we observed a correlation coefficient of 0.329 for ROUGE-1 (F1) and 0.276 for BERTScore (F1), both with a p-value of less than 0.001. The positive monotonic relationship suggests that human evaluation and similarity scores consistently increase together, but not necessarily at a constant rate.

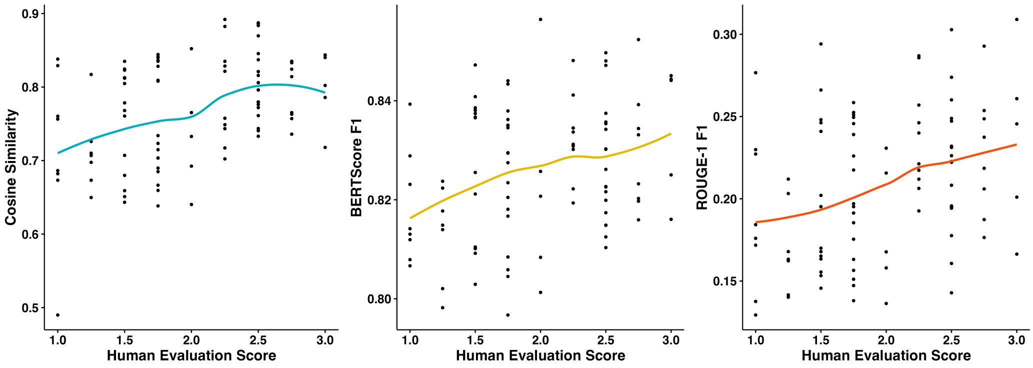

Figure 6: Relationship between human evaluation of topics and their corresponding text similarity scores (cosine similarity, BERTScore F1, and ROUGE-1 F1), computed between original and reconstructed texts. The dataset contains 96 records.

Summaries

Our analysis revealed statistically significant positive monotonic relationships between the text similarity scores for the original and reconstructed texts and four dimensions of human evaluation (Figure 7). As shown in Table 2, the strongest correlations are observed between the reconstruction metrics and the overall quality and coverage dimensions. This pattern is consistent across all three metrics: cosine similarity, BERTScore (F1), and ROUGE-1 (F1). The Spearman's correlation coefficient between the overall quality and the cosine similarity was 0.311 with a p-value of less than 0.0001, while the correlation between coverage and cosine similarity was 0.303. Table 2 presents the Spearman's correlation coefficients between various human evaluation dimensions and text similarity scores between pairs of original and reconstructed texts. Notably, we observe weaker correlations with accuracy and coherence, suggesting that the proposed evaluation method may not fully capture these aspects.

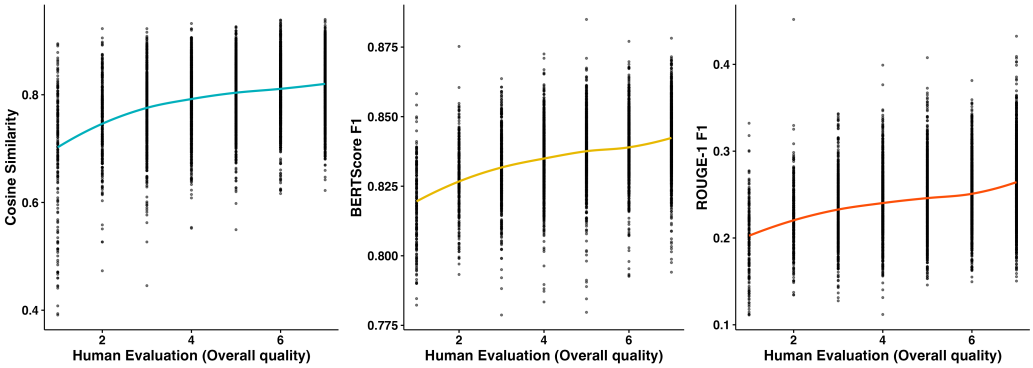

Figure 7: Relationship between human evaluation of summaries (overall quality) and similarity scores (cosine similarity, BERTScore F1, and ROUGE-1 F1), computed between original and reconstructed texts. The dataset contains 5,552 records.

These findings indicate that the proposed evaluation method, particularly when using cosine similarity, aligns with human judgments in evaluating topic names. It also aligns with human judgments regarding the overall quality of summaries and coverage. Here, "coverage" closely mirrors

	ROUGE-1 F1	**BERTScore F1**	**Cosine Similarity**
Overall	0.326297	0.376997	0.311033
Accuracy	0.196744	0.240014	0.149541
Coverage	0.325172	0.368874	0.302683
Coherence	0.110193	0.147968	0.148868

Table 2: Spearman's correlation coefficients between human evaluation dimensions of summaries and similarity scores between original and reconstructed texts.

the concept of "completeness" we considered as a quality metric for topic names, emphasizing the preservation of key information. This alignment validates the potential of our approach for the automated evaluation of various text compression methods in terms of information preservation. It also highlights areas where further refinement may be needed to better capture aspects such as coherence and accuracy, as indicated by the lower correlation observed in the case of summaries. In this context, "coherence" corresponds to "clarity", while "accuracy" corresponds to "correctness" in topic names quality.

6. Limitations

Our approach has some limitations. It is computationally intensive, since it requires the generation of reconstructed texts using LLMs in order to compare them with the original texts. Moreover, LLMs inherently incorporate stochasticity while generating their outputs, meaning that identical prompts may yield slight variations in generated texts across multiple runs. This randomness can introduce variability in reconstructed texts, complicating evaluations that would require several runs or a larger amount of data to obtain reliable results. Additionally, if the same model is used for both compression and reconstruction, the process may benefit from a form of self-bias, potentially leading to artificially better reconstructions. Furthermore, the applied similarity metric (cosine similarity) has inherent limitations, potentially leading to error accumulation.

Despite using recent data to mitigate possible overlap of the testing data with texts that could have been used for LLM training, it is possible that certain selected text categories may align with topics or specific terms that LLMs have encountered before, leading to potential bias. This factor could also impact the generalizability of the approach, particularly in niche domains where familiarity with specific terminology may require domain-adapted LLMs. Additionally, our method focuses on evaluating topic names for individual texts rather than for groups of related texts, such as in traditional topic modeling. Adapting this method for such scenarios would require additional research.

Furthermore, while our evaluation method focuses on measuring information preservation, it does not assess the cognitive load associated with understanding the generated topic names. According to Zipf's principle of least effort, it is essential to evaluate cognitive demands, as even highly informative topics may impose a higher cognitive load, potentially impacting usability and interpretability. Lastly, the human evaluation of topics used in the study relies on judgments from a small group of four experts without evaluating specific topic qualities, limiting the generalizability of these findings. By contrast, the human judgment dataset for summaries includes a broader pool of evaluators, providing a more representative assessment. This factor may affect the robustness of our conclusions regarding the alignment with information preservation.

7. Conclusion and Future Work

This study presents a novel method for automatically evaluating human-written and generated topic names, focusing on information preservation. Our method leverages LLMs to reconstruct the original text from topic names and assesses semantic alignment through text similarity metrics like cosine similarity, BERTScore, and ROUGE. Results of the study indicate that this reconstruction-based evaluation method offers a holistic assessment of topic names that inherently covers such topic name qualities as relevance, completeness, clarity, and correctness and also aligns with human judgment. Additionally, the method is adaptable to various forms of semantic compression, such as summaries, headlines, and keywords, suggesting its potential for broader applications in evaluating compressed representations.

Future research will focus on obtaining more human preference data and explicitly evaluating how the proposed method addresses specific topic qualities (relevance, completeness, clarity, and correctness). Additionally, it will extend this method to groups of texts, where a single topic name represents a collection of related texts. Furthermore, incorporating cognitive load assessments will be a critical step in evaluating the balance between information preservation and ease of understanding. By measuring the cognitive demands required to interpret topic names, future research can optimize topic name quality to maximize informativeness without overburdening users. Finally, studies involving larger and more diverse groups of human evaluators are needed to deepen our understanding of human preferences in topic quality, and they may also reveal variations in expectations based on both the utility of the topics and distinct user personas.

8. Acknowledgment

We thank the anonymous reviewers for their time and effort reviewing our manuscript. Their comments and suggestions have helped us refine our arguments, clarify key points, and enhance the clarity of the manuscript. This research was funded by Flanders Innovation & Entrepreneurship (VLAIO), grant HBC.2021.0222.

References

AlSumait, Loulwah, Daniel Barbará, James Gentle, and Carlotta Domeniconi (2009), Topic significance ranking of lda generative models, *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'09, Springer-Verlag, Berlin, Heidelberg, p. 67–82.

Bartlett, F. C. (1932), *Remembering: A Study in Experimental and Social Psychology*, Cambridge University Press, Cambridge.

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin (2017), An automatic approach for document-level topic model evaluation, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Association for Computational Linguistics, Vancouver, Canada, pp. 206–215. https://aclanthology.org/K17-1022.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), Latent dirichlet allocation, *J. Mach. Learn. Res.* **3** (null), pp. 993–1022, JMLR.org.

Boyd-Graber, Jordan, Yuening Hu, and David Mimno (2017), Applications of topic models, *Foundations and Trends® in Information Retrieval* **11** (2-3), pp. 143–296. http://dx.doi.org/10.1561/1500000030.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei (2009), Reading tea leaves: How humans interpret topic models, *in* Bengio, Y., D. Schuurmans, J. Lafferty,

C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, Vol. 22, Curran Associates, Inc.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, ..., and Jason Wei (2022), Scaling instruction-finetuned language models. https://arxiv.org/abs/2210.11416.

Churchill, Rob and Lisa Singh (2022), The evolution of topic modeling, *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3507900.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei (2020), Topic modeling in embedding spaces, *Transactions of the Association for Computational Linguistics* **8**, pp. 439–453, MIT Press, Cambridge, MA. https://aclanthology.org/2020.tacl-1.29/.

Gilbert, Henry, Michael Sandborn, Douglas C. Schmidt, Jesse Spencer-Smith, and Jules White (2023), Semantic compression with large language models. https://arxiv.org/abs/2304.12512.

Grice, Paul (1975), Logic and conversation, *in* Cole, P. and J. Morgan, editors, *Syntax and Semantics*, Vol. 3 of *Speech Acts*, Academic Press, New York, pp. 41–58.

Grootendorst, Maarten (2020), KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265.

Grusky, Max, Mor Naaman, and Yoav Artzi (2018), Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 708–719.

Hanna, Michael and Ondřej Bojar (2021), A fine-grained analysis of BERTScore, *Proceedings of the Sixth Conference on Machine Translation*, Association for Computational Linguistics, Online, pp. 507–517. https://aclanthology.org/2021.wmt-1.59/.

Hoyle, Alexander, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik (2021), Is automated topic model evaluation broken? the incoherence of coherence, *in* Ranzato, M., A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc., pp. 2018–2033.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, ..., and Ting Liu (2024), A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Trans. Inf. Syst.*, Association for Computing Machinery, New York, NY, USA. Just Accepted. https://doi.org/10.1145/3703155.

Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, ..., and William El Sayed (2024), Mixtral of Experts. https://arxiv.org/abs/2401.04088.

Kapoor, Satya, Alex Gil, Sreyoshi Bhaduri, Anshul Mittal, and Rutu Mulkar (2024), Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling. https://arxiv.org/abs/2409.15626.

Kirilenko, Andrei and Svetlana Stepchenkova (2024), Automated topic analysis with large language models, *in* Berezina, Katerina, Lyndon Nixon, and Aarni Tuomi, editors, *Information and Communication Technologies in Tourism 2024*, Springer Nature Switzerland, Cham, pp. 29–34.

Korencic, Damir, Strahil Ristov, Jelena Repar, and Jan Snajder (2021), A topic coverage approach to evaluation of topic models, *IEEE Access* **9**, pp. 123280–123312, Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109

Kosar, Andriy, Guy De Pauw, and Walter Daelemans (2024), Comparative Evaluation of Topic Detection: Humans vs. LLMs, *Computational Linguistics in the Netherlands Journal* **13**, pp. 91–120. https://www.clinjournal.org/clinj/article/view/173.

Lau, Jey Han, David Newman, and Timothy Baldwin (2014), Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, *in* Wintner, Shuly, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Gothenburg, Sweden, pp. 530–539. https://aclanthology.org/E14-1056.

Li, Haitao, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu (2024), Llms-as-judges: A comprehensive survey on llm-based evaluation methods. https://arxiv.org/abs/2412.05579.

Lin, Chin-Yew (2004), ROUGE: A package for automatic evaluation of summaries, *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, pp. 74–81. https://aclanthology.org/W04-1013/.

Lund, Jeffrey, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi (2019), Automatic evaluation of local topic quality, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 788–796. https://aclanthology.org/P19-1076.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011), Optimizing semantic coherence in topic models, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 262–272. https://aclanthology.org/D11-1024.

Mu, Yida, Chun Dong, Kalina Bontcheva, and Xingyi Song (2024), Large language models offer an alternative to the traditional approach of topic modelling, *in* Calzolari, Nicoletta, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, Torino, Italia, pp. 10160–10171.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin (2010), Automatic evaluation of topic coherence, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Association for Computational Linguistics, USA, p. 100–108.

OpenAI (2023), GPT-4 Technical Report. https://arxiv.org/abs/2303.08774.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, ..., and Ryan Lowe (2022), Training language models to follow instructions with human feedback. https://arxiv.org/abs/2203.02155.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020), Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, *Journal of Machine Learning Research* **21** (140), pp. 1–67. http://jmlr.org/papers/v21/20-074.html.

Shannon, C. E. (1948), A mathematical theory of communication, *The Bell System Technical Journal* **27**, pp. 379–423, 623–656.

Stammbach, Dominik, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash (2023), Re-visiting Automated Topic Model Evaluation with Large Language Models. https://arxiv.org/abs/2305.12152.

Stiennon, Nisan, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano (2022), Learning to summarize from human feedback. https://arxiv.org/abs/2009.01325.

Wang, Jiapeng and Yihong Dong (2020), Measurement of Text Similarity: A Survey, *Information*. https://www.mdpi.com/2078-2489/11/9/421.

Xiao, Shitao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie (2024), C-pack: Packed resources for general chinese embeddings. https://arxiv.org/abs/2309.07597.

Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, ..., and Zhihao Fan (2024), Qwen2 technical report. https://arxiv.org/abs/2407.10671.

Yu, Haizi and Lav R. Varshney (2024), Semantic compression with information lattice learning. https://arxiv.org/abs/2404.03131.

Zhang, Ming, Chengzhang Li, Meilin Wan, Xuejun Zhang, and Qingwei Zhao (2024), ROUGE-SEM: Better evaluation of summarization using ROUGE combined with semantics, *Expert Systems with Applications* **237**, pp. 121364. https://www.sciencedirect.com/science/article/pii/S0957417423018663.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2020), BERTScore: Evaluating text generation with BERT, *International Conference on Learning Representations*.

Zhou, Kaitlyn, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky (2022), Problems with cosine as a measure of embedding similarity for high frequency words, *in* Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Dublin, Ireland, pp. 401–423. https://aclanthology.org/2022.acl-short.45/.

Zipf, George Kingsley (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley, Cambridge.

Appendix A. Evaluating Topic Names via Text Reconstruction

A.1 Compression and Reconstruction Analysis

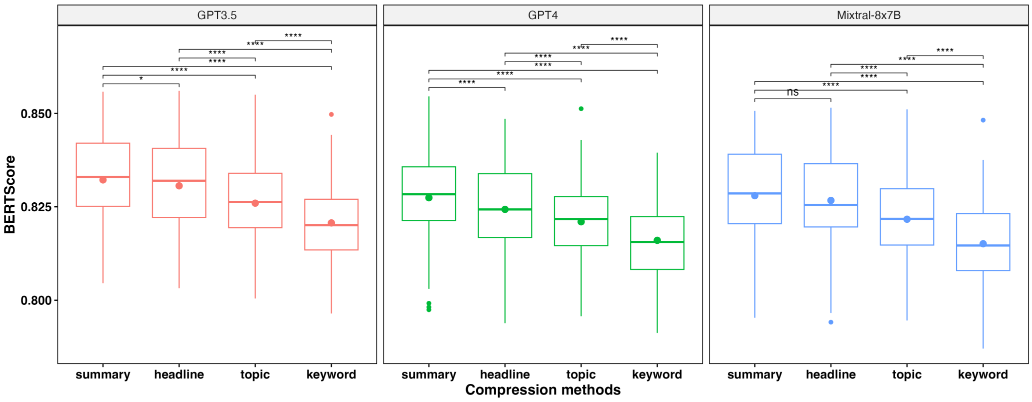

Figure 8: Side-by-side box plots of BERTScore F1 in four compression methods, split by three reconstruction models. Each compression method group contains 100 records, with each record representing the median metric score for compression models within that method. Statistically significant mean differences between methods pairs are denoted by asterisks. Method: two-way repeated measures ANOVA, followed by pairwise paired t-test comparisons with Benjamini–Hochberg correction.

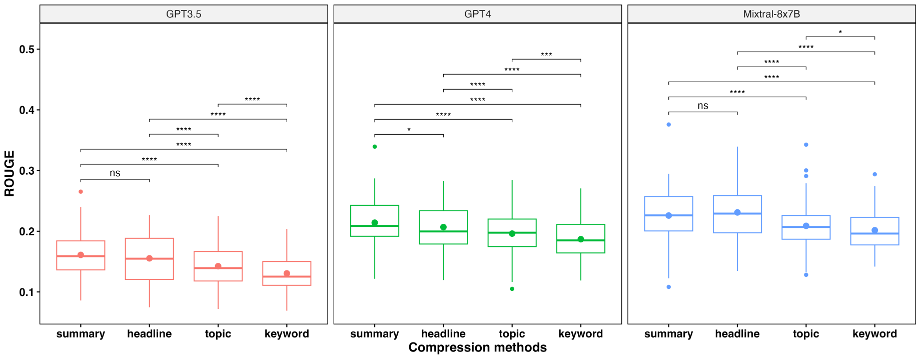

Figure 9: Side-by-side box plots of ROUGE-1 F1 in four compression methods, split by three reconstruction models. Each compression method group contains 100 records, with each record representing the median metric score for compression models within that method. Statistically significant mean differences between methods pairs are denoted by asterisks. Method: two-way repeated measures ANOVA, followed by pairwise paired t-test comparisons with Benjamini–Hochberg correction.

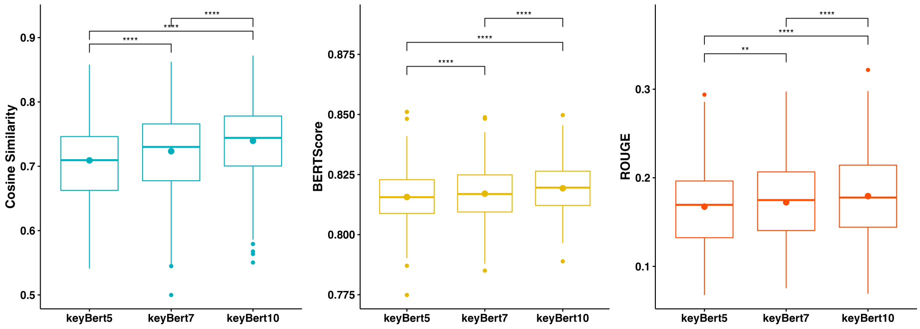

Figure 10: Side-by-side box plots compare the distributions of three metrics (cosine similarity, BERTScore F1, and ROUGE-1 F1) across different numbers of keywords (5, 7, 10 keywords, respectively) in the KeyBERT model. For each metric, the mean differences between all groups are statistically significant. Method: two-way repeated measures ANOVA, followed by pairwise paired t-tests with Benjamini–Hochberg correction.

	GPT3.5		GPT4		Mixtral-8x7B	
Compression model	mean	sd	mean	sd	mean	sd
GPT3.5	0.738	0.079	0.759	0.073	0.758	0.068
GPT4	0.753	0.067	0.767	0.065	0.768	0.065
GPT3.5FT	0.758	0.073	0.772	0.063	0.771	0.060
FlanT5FT	0.766	0.068	0.780	0.066	0.774	0.066

Table 3: Cosine similarity results for topics: compression models in rows and reconstruction models in columns.

A.2 Prompt, Compression, and Reconstruction Examples

A.2.1 Prompt Example

System Message

You are assigned the role of a human expert tasked with reconstructing news text from provided information. Each task will supply you with one of the following: keywords, a topic, a headline, or a summary. Your objective is to analyze these elements and create a news article that closely represents the original content in its entirety within a maximum of 1000 words.

User Message

Here is the given information from which to reconstruct:

`{INPUT_TEXT}`

A.2.2 Examples of Semantic Compression

Original text:

The first Black president of Harvard, who resigned on Tuesday after a successful rightwing campaign to oust her, warned that the tactics used against her were “merely a single skirmish in a broader war to unravel public faith in pillars of American society”. “Trusted institutions of all types

	GPT3.5		GPT4		Mixtral-8x7B	
Compression model	mean	sd	mean	sd	mean	sd
GPT3.5	0.789	0.064	0.802	0.059	0.794	0.064
GPT4	0.817	0.059	0.824	0.052	0.807	0.059
FlanT5FT	0.808	0.056	0.816	0.056	0.809	0.063

Table 4: Cosine similarity results for summaries: compression models in rows and reconstruction models in columns.

– from public health agencies to news organizations – will continue to fall victim to coordinated attempts to undermine their legitimacy and ruin their leaders' credibility," Claudine Gay wrote in the New York Times on Wednesday, a day after she announced she was resigning from her position and returning to her role as a faculty member. Gay, a political scientist who was appointed six months ago as the first Black woman to serve as president of Harvard, had the shortest tenure in the university's 388-year history. In her first major comment since her official Harvard statement announcing her resignation, Gay admitted bluntly that "I made mistakes". But she also argued that her invitation to testify to Congress about antisemitism on elite college campuses had been "a well-laid trap" and that "the campaign against me was about more than one university and one leader". The playbook that had successfully been used against her would soon be mobilized against other institutional leaders, Gay warned. "For the opportunists driving cynicism about our institutions, no single victory or toppled leader exhausts their zeal," she said. For the past month, the campaign against Gay, which included prominent Harvard donors, had centered on allegations of antisemitism and plagiarism in her academic work, focusing on her widely criticised comments during a December congressional hearing on antisemitism on college campuses, and on multiple passages in her academic work that closely resembled the work of other scholars, without the appropriate citations. Gay once again said she regretted not speaking forcefully enough against antisemitism on campus in congressional remarks that had sparked bipartisan backlash, writing: "I should have stated more forcefully what all people of good conscience know: Hamas is a terrorist organization that seeks to eradicate the Jewish state" and that she had "neglected to clearly articulate that calls for the genocide of Jewish people are abhorrent and unacceptable and that I would use every tool at my disposal to protect students from that kind of hate". [...]

Original headline:
Ousted Harvard president Claudine Gay warns of 'a broader war' in op-ed.

Summaries:
- **GPT4:** Claudine Gay resigns as Harvard's president amid rightwing campaign, citing broader war against public trust in institutions and facing allegations of antisemitism and plagiarism.
- **GPT3.5:** Harvard's first Black president resigns amid rightwing campaign, warns of broader attacks on institutions and leaders.
- **T5FT:** Claudine Gay warns of coordinated attempts to undermine legitimacy and ruin credibility of trusted institutions.

Topics[9]**:**
- **GPT4:** Harvard President Claudine Gay's Resignation (0.201, 0.820, 0.794)
- **GPT3.5:** Harvard's First Black President Resignation (0.215, 0.819, 0.754)
- **GPT3.5FT:** Harvard President's Resignation and Racism (0.196, 0.821, 0.741)
- **FLANT5FT:** Claudine Gay's Harvard Resignation (0.189, 0.815, 0.751)

---

[9] The three scores represent ROUGE-1 F1, BERTscore F1, and cosine similarity between the original and reconstructed text. Reconstruction performed by GPT-4.

Keywords:

- **KeyBERT5:** harvard, tenure, resigned, academics, academic
- **KeyBERT7:** harvard, tenure, resigned, academics, academic, resigning, president
- **KeyBERT10:** harvard, tenure, resigned, academics, academic, resigning, president, allegations, faculty, bipartisan

A.2.3 Examples of Semantic Reconstruction

Original headline:

Ousted Harvard president Claudine Gay warns of 'a broader war' in op-ed.

Reconstructed text (GPT4):

In a compelling op-ed, Claudine Gay, the recently ousted president of Harvard University, has issued a stark warning about what she perceives as a 'broader war' affecting higher education and society at large. Gay, who had served in her role at one of the world's most prestigious universities, did not specify the exact nature of this war in her commentary. However, her words suggest a deep concern for the current state and future direction of academia and its role in the broader societal context.

Gay's tenure at Harvard was marked by her efforts to expand access to education, promote diversity and inclusion, and foster a culture of innovation and excellence. Despite these efforts, her presidency came to an abrupt end, a decision that has sparked considerable debate and speculation within academic circles and beyond. The reasons behind her ousting have not been made public, leading to further speculation about the challenges facing leaders in higher education. [...]

Summary (GPT4):

Claudine Gay resigns as Harvard's president amid rightwing campaign, citing broader war against public trust in institutions and facing allegations of antisemitism and plagiarism.

Reconstructed text (GPT3.5):

Claudine Gay, the president of Harvard University, has announced her resignation amidst a right-wing campaign targeting her. In her statement, Gay cited a broader war against public trust in institutions as one of the reasons for her decision to step down. The resignation comes as Gay faced allegations of antisemitism and plagiarism, further complicating her tenure at the prestigious university. Gay, who became the first Black president of Harvard in 2021, has been a target of criticism from conservative groups and media outlets. The allegations of antisemitism and plagiarism have added fuel to the fire, leading to increased pressure on Gay to resign. Despite the challenges she faced during her time as president, Gay expressed gratitude for the opportunity to lead Harvard and emphasized the importance of upholding trust in institutions. [...]

Topic (FLANT5FT):

Claudine Gay's Harvard Resignation.

Reconstructed text (Mixtral-8x7B):

Claudine Gay, the first woman of color to lead Harvard College, has announced her resignation from the prestigious institution. Gayś departure comes after four years of service as the Dean of Harvard College, during which she made significant contributions to the universityś academic and social life.

Gay, who is also a professor of Government and African and African American Studies, was appointed as the Dean of Harvard College in 2018. She succeeded Rakesh Khurana, who is now the Dean of Harvard Business School. During her tenure, Gay focused on enhancing the student experience, promoting diversity and inclusion, and strengthening the collegeś academic programs. [...]

Keywords (KeyBERT 7):

harvard, tenure, resigned, academics, academic, resigning, president

Reconstructed text (GPT3.5):

Harvard University’s president has resigned after a tenure marked by significant academic achievements and controversies. The decision to step down was announced by the prestigious institution, citing personal reasons for the departure. The outgoing president, known for his dedication to academic excellence, expressed gratitude for the opportunity to lead the university. During his tenure, Harvard saw advancements in various academic fields and research initiatives, solidifying its position as a top academic institution globally. The resignation has sparked discussions among academics and the broader community about the future leadership of the university and the impact of the president’s departure on ongoing projects and initiatives. Harvard’s search for a new president is expected to attract top academic talent from around the world, as the institution seeks to maintain its reputation for excellence in education and research.

A.2.4 Examples of Generated Topics with Evaluation Scores

Text ID	Topic Source	Text	ROUGE	BERTScore	Cosine
Sport					
3	FLANT5FT	NFL Playoff Wins and Investments	0.1398	0.8026	0.6615
3	GPT3.5	Sports Predictions for 2024	0.1850	0.8111	0.6420
3	GPT4	Sports Predictions and Developments	0.1326	0.8058	0.5954
3	GPT3.5FT	2024 Sports Predictions and Developments	0.1438	0.8034	0.5913
9	FLANT5FT	Wimbledon Park Expansion Battle	0.2678	0.8393	0.8355
9	GPT3.5	Wimbledon Park expansion dispute	0.2556	0.8359	0.8305
9	GPT4	Wimbledon Expansion Controversy	0.2140	0.8311	0.8230
9	GPT3.5FT	Wimbledon Park Expansion Controversy	0.2336	0.8309	0.7825
Culture					
17	FLANT5FT	Dave Chappelle’s ’The Dreamer’ Special	0.1370	0.8173	0.8258
17	GPT4	Criticizing Chappelle’s Transphobic Comedy	0.1454	0.8075	0.7295
17	GPT3.5FT	Critique of Chappelle’s Comedy Special	0.1354	0.8067	0.7179
17	GPT3.5	Controversial Comedy Specials Criticism	0.1269	0.8017	0.6343
19	GPT3.5	Tasmania’s political sculpture discovery	0.1699	0.8212	0.7857
19	FLANT5FT	Tasmania’s Historic Political Statue Discovery	0.1565	0.8194	0.7852
19	GPT4	Tasmania’s satirical protest statue discovery	0.1402	0.8073	0.7439
19	GPT3.5FT	Tasmanian Political Protest Sculpture	0.1632	0.8083	0.7182
Environment					
26	GPT3.5	Polynesian Canoe Hōkūle‘a Cultural Renaissance	0.1982	0.8401	0.8371
26	GPT3.5FT	Hōkūle‘a’s Traditional Navigation Revival	0.1972	0.8448	0.8351
26	GPT4	Hōkūle‘a’s Traditional Navigation Voyage	0.2107	0.8374	0.8324
26	FLANT5FT	Hawaiian Sailing Canoe Celebration	0.1715	0.8134	0.7466
29	FLANT5FT	Azerbaijan’s Cop29 Climate Summit Committee	0.1802	0.8270	0.8183
29	GPT4	Cop29 Committee Gender Imbalance	0.2017	0.8263	0.7645
29	GPT3.5	Gender Disparity in COP29 Summit	0.1988	0.8231	0.7467
29	GPT3.5FT	Gender Imbalance in Cop29 Summit	0.1929	0.8223	0.7420
Food					
38	GPT4	Food Mill Appreciation and Recipe	0.2590	0.8205	0.7907
38	GPT3.5FT	Food Mill’s Culinary Uses	0.2212	0.8354	0.7731
38	GPT3.5	Versatility of Food Mill in Cooking	0.1935	0.8284	0.7489
38	FLANT5FT	Victor Simon’s Food Mill Patent	0.1409	0.8183	0.7100
40	GPT3.5	Wine value and recommendations	0.2293	0.8174	0.8104
40	GPT3.5FT	Navigating Wine Value and Quality	0.1982	0.8155	0.8020
40	FLANT5FT	Wine Price and Quality Speculation	0.1800	0.8176	0.7872
40	GPT4	Wine Value and Pricing	0.1937	0.8216	0.7761

Table 5: Examples of randomly selected generated topics from 100 texts across 10 news sections, along with corresponding text similarity scores (ROUGE-1 F1, BERTScore F1, cosine similarity) between the original and reconstructed texts. Text reconstruction is performed by GPT-4 model. Texts are sorted by cosine similarity (Part 1 of 2).

Text ID	Topic Source	Text	ROUGE	BERTScore	Cosine
Business					
43	GPT3.5FT	Morrisons' Customer-Centric Turnaround Strategy	0.1911	0.8263	0.7781
43	FLANT5FT	Morrisons' Restructuring Strategy	0.1802	0.8200	0.7615
43	GPT4	Morrisons' Revitalization Strategy	0.1892	0.8283	0.7575
43	GPT3.5	Morrisons supermarket revitalization strategy	0.1707	0.8221	0.7557
44	FLANT5FT	Thames Water's Service Commitment Plan	0.1894	0.8260	0.7939
44	GPT4	Thames Water's Regulatory Update	0.1856	0.8259	0.7783
44	GPT3.5FT	Thames Water's Regulatory Order	0.2004	0.8304	0.7736
44	GPT3.5	Thames Water service improvement	0.2105	0.8273	0.7454
World news					
57	GPT3.5FT	West Bank Militant Tunnels	0.1853	0.8259	0.7489
57	FLANT5FT	West Bank Refugee Camp Tunnels	0.1802	0.8225	0.7483
57	GPT3.5	West Bank IDF raids escalation	0.1688	0.8262	0.7252
57	GPT4	West Bank violence and Israeli raids	0.1669	0.8206	0.7224
59	GPT3.5	Israeli incitement to genocide	0.1787	0.8106	0.7735
59	FLANT5FT	Israelis Demand Legal Action on Gaza Incitement	0.1887	0.8172	0.7659
59	GPT3.5FT	Israeli Incitement to Genocide Accusations	0.1592	0.8103	0.7580
59	GPT4	Israeli Incitement to Genocide Concerns	0.1777	0.8133	0.7306
Politics					
65	FLANT5FT	Kemi Badenoch's Tory Plotting	0.1580	0.8251	0.7570
65	GPT4	Badenoch's potential Tory leadership plot	0.1575	0.8246	0.7539
65	GPT3.5FT	Kemi Badenoch's Alleged Leadership Ambitions	0.1684	0.8174	0.7477
65	GPT3.5	Conservative Party Leadership Dynamics	0.1530	0.8013	0.5610
68	GPT3.5FT	UK Political Missteps and Scandals	0.1335	0.8028	0.6992
68	GPT4	UK Politicians' Detachment Critique	0.1484	0.8007	0.6551
68	GPT3.5	Politicians' Conduct Critique	0.1259	0.7991	0.6166
68	FLANT5FT	Political Inquiry on Covid Inquiry	0.1000	0.7977	0.5560
Music					
77	GPT3.5	Black Grape music interview	0.1458	0.8130	0.7130
77	GPT3.5FT	Black Grape's Album and Reunion	0.1450	0.8156	0.6836
77	GPT4	Black Grape's New Album Release	0.1156	0.8108	0.6695
77	FLANT5FT	Black Grape's New Album and Death	0.1562	0.8128	0.6685
80	GPT4	Singer-songwriter Melanie's death	0.2491	0.8281	0.8844
80	FLANT5FT	Melanie Safka's Death and Legacy	0.2373	0.8299	0.8744
80	GPT3.5	Melanie Safka's Death	0.2203	0.8244	0.8651
80	GPT3.5FT	Singer Melanie Safka's Death	0.2180	0.8247	0.8468
Society					
82	GPT4	London Mould Crisis in Social Housing	0.1869	0.8231	0.8280
82	GPT3.5	Mould in UK social housing	0.1791	0.8224	0.8174
82	FLANT5FT	UK Mold Concerns	0.1966	0.8197	0.8142
82	GPT3.5FT	Social Housing Mould Crisis	0.1740	0.8241	0.7910
89	GPT3.5FT	Debate Over Physician Associates in NHS	0.1975	0.8320	0.7796
89	FLANT5FT	NHS's Physician Associate Expansion Critique	0.2238	0.8287	0.7777
89	GPT3.5	Physician Associates in NHS	0.2084	0.8333	0.7715
89	GPT4	NHS Physician Associates Controversy	0.2095	0.8304	0.7679
Education					
94	GPT3.5FT	Oliver's Ofsted Chief Inspector Role	0.1821	0.8203	0.7915
94	GPT4	Ofsted's New Chief Inspector Appointment	0.1678	0.8150	0.7790
94	FLANT5FT	Ofsted Chief Inspector's Career Path	0.1761	0.8139	0.7323
94	GPT3.5	School Leadership and Improvement in England	0.1588	0.8158	0.7232
95	FLANT5FT	Disadvantaged Parents' Stem Careers	0.2309	0.8271	0.8161
95	GPT3.5FT	Disadvantaged Parents' Stem Career Concerns	0.2500	0.8246	0.8042
95	GPT4	Stem Career Access Inequality	0.2088	0.8149	0.7273
95	GPT3.5	Barriers to STEM Careers	0.1954	0.8144	0.7079

Table 6: Examples of randomly selected generated topics from 100 texts across 10 news sections, along with corresponding text similarity scores (ROUGE-1 F1, BERTScore F1, cosine similarity) between the original and reconstructed texts. Text reconstruction is performed by GPT-4 model. Texts are sorted by cosine similarity (Part 2 of 2).